REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY) 16-08-2010	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 1-Sep-2009 - 31-May-2010

4. TITLE AND SUBTITLE Inference for Identity Management -- Final Report	5a. CONTRACT NUMBER W911NF-09-1-0468
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 611102
6. AUTHORS Carlo Tomasi	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES Duke University Office of Research Support Duke University Durham, NC 27705 -	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	10. SPONSOR/MONITOR'S ACRONYM(S) ARO 11. SPONSOR/MONITOR'S REPORT NUMBER(S) 56998-MA-II.1

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

A computational framework has been developed to carry out identity management, that is, the automatic inference of the identities of targets tracked by surveillance systems that cover wide areas such as a shopping mall or a large harbor. People or vehicles may remain invisible to the system for long periods of time as they move between sensors. Identity management attempts to infer from uncertain measurements who or what is where at all times.

15. SUBJECT TERMS

camera networks, probabilistic inference, computer vision, identity management

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Carlo Tomasi
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU	UU		19b. TELEPHONE NUMBER 919-660-6539

Report Title

Inference for Identity Management -- Final Report

ABSTRACT

A computational framework has been developed to carry out identity management, that is, the automatic inference of the identities of targets tracked by surveillance systems that cover wide areas such as a shopping mall or a large harbor. People or vehicles may remain invisible to the system for long periods of time as they move between sensors. Identity management attempts to infer from uncertain measurements who or what is where at all times.

The following work was performed in this short-term project:

Fleshed out and streamlined the mathematical framework for identity management. This required significant changes at the core of the framework, and several of the ideas built on top of this had to be adapted or reinvented as well, prompting a systematic reformulation of the mathematics.

Studied and tested algorithms from the literature to be used, either directly or in modified form, in the core inference engine of an identity management system.

Developed a computationally efficient method for finding high-likelihood identity assignments given a graph of association probabilities between sensor observations. This method efficiently solves the batch version of the main estimation problem underlying identity management.

List of papers submitted or published that acknowledge ARO support during this reporting period. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Number of Papers published in peer-reviewed journals: 0.00

(b) Papers published in non-peer-reviewed journals or in conference proceedings (N/A for none)

Number of Papers published in non peer-reviewed journals: 0.00

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts): 0

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts): 0

(d) Manuscripts

Number of Manuscripts: 0.00

Patents Submitted

Patents Awarded

Graduate Students

NAME	PERCENT_SUPPORTED
Zhiqiang Gu	0.50
FTE Equivalent:	**0.50**
Total Number:	**1**

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Carlo Tomasi	0.11	No
FTE Equivalent:	**0.11**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Inference for Identity Management

FINAL REPORT

Carlo Tomasi — Duke University

tomasi@cs.duke.edu

Short-Term Innovative Research Grant
BAA W911NF-07-R-0003-03
US Army RDECOM Acquisition Center

Abstract

A computational framework has been developed to carry out *identity management*, that is, the automatic inference of the identities of targets tracked by surveillance systems that cover wide areas such as a shopping mall or a large harbor. People or vehicles may remain invisible to the system for long periods of time as they move between sensors. Identity management attempts to infer from uncertain measurements who or what is where at all times.

The following work was performed in this short-term project:

- Fleshed out and streamlined the mathematical framework for identity management. This required significant changes at the core of the framework, and several of the ideas built on top of this had to be adapted or reinvented as well, prompting a systematic reformulation of the mathematics.
- Studied and tested algorithms from the literature to be used, either directly or in modified form, in the core inference engine of an identity management system.
- Developed a computationally efficient method for finding high-likelihood identity assignments given a graph of association probabilities between sensor observations. This method efficiently solves the batch version of the main estimation problem underlying identity management.

Contents

Illustration Captions and Appendices

Figure 1: Algorithms that measure the cost of a cut by the sum of its edge weights will not remove the weak, spurious edges e_1, e_2 between components A and B of this graph until after "dangling" connections like e_3, e_4 are severed. Page 6

Figure 2: Top left: An observation graph. Blue edges correspond to strong associations, red edges to weak ones. Other panels: Partitions in the complete filtration for the graph at top left. The ranges of π that result in each partition are shown above each partition. The boxed partition is the most persistent one. Page 9

Appendix A: Bayesian Filtering Page 10

1 Statement of the Problem Studied

Surveillance systems often cover wide areas such as a shopping mall, a city's subway system, a large harbor, or several city blocks around a sensitive installation. People or vehicles (henceforth *targets*) are detected by sensors (typically cameras or camera/LIDAR combinations) in different locations and at different points in time. Targets may remain invisible for several seconds or minutes as they move through areas with no coverage. Even when they remain in the field of view of, say, a single camera, targets may be occluded by (or become indistinguishable from) other targets, or behind objects. Even when a target is distinctly visible, its appearance may vary as a result of changes in lighting, distance, pose, or sensing parameters.

The fundamental challenge then arises of how to infer an accurate, consistent picture of the world from intermittent, uncertain observations. This problem has been called *identity management* in the literature [4], and has been the main focus of the work performed under this short-term grant.

2 Summary of the Most Important Results

The mathematical framework sketched out in the grant proposal was fleshed out and streamlined, as described in Section 2.1 below. This required significant changes at the core of the framework, and several of the ideas built on top of this had to be adapted or reinvented as well, prompting a systematic reformulation of the mathematics. The most promising algorithms from the literature, discussed in Section 2.2, were then tested empirically with simulated data, as shown in section 2.3. Dissatisfaction with these results were tied to the so-called *additivity problem*, described in Section 2.4. This difficulty prompted the development of a new formulation of the batch version of the main inference problem for identity management, the *partition filtration* associated with a graph of observations. The partition filtration is introduced in Section 2.5.

2.1 The Mathematical Framework

An important improvement in the mathematical framework sketched in the proposal for this grant is a shift in the basic representation of observations and states. In the old version, a combinatorial structure called a multipartite partition was used to describe deterministic knowledge, or belief, as to which set of observations relate to the same target. Specifically, the set of all observations can be partitioned into sets, one set per target. This partition was made multipartite, in the proposal, to capture the notion that for certain pairs of observations it is possible to know withe certainty that they refer to distinct targets. For instance, a single sensor is assumed to be able to produce at most one observation of a target at any point in time. So two observations produced simultaneously by the same sensor are assumed to relate to different target. Another example is a pair of observations that, although taken at different points in time, come from sensors that are so distant that no target could have moved from one sensor to the other in the given time interval. More generally, the observations made by a set of sensors in a surveillance network can be split into K sets such that the sightings within a set are certain to refer to distinct identities. This split makes the partitions multipartite.

Initially in this effort, mathematics and simulations were all carried out in this context of multipartite partions. After some time, however, it became clear that constraining the partitions to be multipartite led to more difficulties than it addressed. The main reason for this is that, while the partition aspect of a multipartite partition captures equality – as in "observation a is equal to observation b" – , its multipartite aspect captures non-equality – "a is not equal to b." Since the notion of "not equal" is not transitive, the resulting clash produced increasing complications both in data structure and inference methods.

The solution to this difficulty turned out to be both simple and effective. Multipartition is removed, and non-equality is captured by (i) assigning a zero probability to pairs of observations that are known with certainty to refer to distinct targets, and (ii) *clamping* these zero values – that is, forcing them to remain unchanged – during inference. This change in the representation that lives at the very core of the framework required a systematic modification of everything built on top of it, including changing weighted *match* graphs into weighed *association* graphs – the former requiring multipartition, the latter not.

Rather than describing in detail the history of all the changes, their net result, that is, the new framework, is briefly outlined next. This description follows closely the reasoning in a new grant proposal submitted to ARO for a full-fledged investigation of identity management.

A key representational decision is to capture the information contained in a set of n measurements through m *association* values $0 \leq p_{ij} \leq 1$, defined as the probability that measurements i and j were generated by the same target, given the values of the measurements. This choice contrasts with the more usual approach of summarizing measurements through n feature vectors, one per measurement, and then defining a metric in their space.

Pairwise associations are preferred over individual feature vectors for two reasons. First, sensors used in different parts of the space under observation can be of different types (cameras, LIDARs, proximity sensors, other). The corresponding output values are then heterogeneous, and no single space is likely to fit both types of outputs naturally. Second, the time elapsed *between* the two measurements is an important source of information for computing the association value between them. This computation must consider the distance between the sensors, estimates of travel speeds, and the presence of possible delays (in an airport, delays could come from stores, restaurants, security lines, ...) or accelerators (moving walkways, escalators, ...) between the two measurement stations. These considerations cannot be captured by either measurement alone, but can be incorporated in the association between sensor measurements at different stations.

Because of these reasons, associations are more flexible and potentially richer than separate measurement features. Of course, associations can be computed from metrics defined in a feature space whenever the situation warrants – that is, when measurements happen to be homogeneous. Thus, associations subsume the standard approach, and provide a representational foundation for a broader set of circumstances.

Both the number n of measurements and the number m of association values are growing functions of time t. The $n(t)$ measurements can be represented as the set $V(t)$ of nodes in a growing, weighted *association graph* $G(t) = (V(t), E(t), P(t))$. Two measurements are connected by an edge in $E(t)$ if an association value $p_{ij} \in P(t)$ is available for them, and p_{ij} is the *weight* for that edge.[1] The graph $G(t)$ is generally not complete, as it does not always

[1] For convenience, the reflexive property is ignored throughout this proposal. In other words, a measurement is not associated with – or considered equivalent to – itself.

make sense to establish associations for two measurements. For instance, two sensors may be cameras that look at people from different directions (perhaps from the front and from the back), or across excessively long time intervals, and this may make matching between these views meaningless.

Bayesian estimation is made possible by the second, key contribution made in our work: A method for defining a probability distribution over association graphs – and therefore equivalence graphs as well. This is a new definition, significantly different from what was in our original proposal. Briefly, and following the spirit of the idea that Mallows [6] introduced for permutations, we define a measure of *compatibility* between a partition graph $\Gamma = (V, E, P)$ and and association graph $H = (V, F, Q)$ on the same set of nodes V. Recall that a partition graph is equivalent to a partition of its nodes, so that E is the complete set of edges, P is binary (that is, p_{ij} is either 0 or 1), and specifies a set of disjoint cliques. The compatibility between Γ and H is defined as follows:

$$d(\Gamma, H) = \sum_{E \cap F} |p_{ij} - q_{ij}| \; . \tag{1}$$

This is not a distance. For instance, if the edge set F in H is empty, then $d(\Gamma, H) = 0$ regardless of Γ. More generally, this measure of compatibility does not penalize unspecified edges in the association graph H. As a consequence, ignorance as to whether two observations do or do not correspond to the same identity is compatible with any assignment of identities to observations.

Ignorance, on the other hand, is the basis for the following definition of *dispersion*, a measure of the uncertainty implied by an association graph $H = (V, F, Q)$:

$$\sigma = \sum_{F} \min(q_{ij}, 1 - q_{ij}) + \frac{1}{2} \left[\binom{|V|}{2} - |F| \right] \; . \tag{2}$$

The first term in this measure penalizes uncertain associations, that is values of q_{ij} that are different from either 0 or 1. The second term penalizes missing associations, since the expression in square brackets is the number of edges that are in F and not in the complete graph on V. Each such missing association receives a penalty of $1/2$. The maximum penalty for uncertain associations is also $1/2$, so that the maximum possible dispersion for an association graph with $|V|$ nodes is

$$\frac{1}{2} \binom{|V|}{2} \; ,$$

and is incurred by the empty association graph.

Then, given a distinguished set of C equivalence graphs $\Gamma_1, \ldots, \Gamma_C$ and C dispersion parameters $\sigma_1, \ldots, \sigma_C$, we define the following parametric probability measure on the set A of association graphs:

$$p(G) = e^{-\psi(\boldsymbol{\sigma}, \boldsymbol{\gamma})} \sum_{c=1}^{C} e^{-\frac{d(G, \Gamma_c)}{\sigma_c}} \quad \text{where} \quad \psi(\boldsymbol{\sigma}, \boldsymbol{\gamma}) = \log \sum_{G' \in A} \sum_{c=1}^{C} e^{-\frac{d(G', \Gamma_c)}{\sigma_c}} \tag{3}$$

is the *cumulant* function and

$$\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_C)^T \quad \text{and} \quad \boldsymbol{\gamma} = (\Gamma_1, \ldots, \Gamma_C)^T \; .$$

This probability measure on the set of association graphs provides a conceptual foundation for learning and inference in the context of identity management, based on the general principles of Sequential Importance Sampling for Bayesian filtering [1].

Specifically, given a sequence $\mathbf{y}(t_0), \mathbf{y}(t_1), \ldots$ of measurements over time, in the form of association values p_{ij}, the current belief about the true state $\mathbf{x}(t)$ – a particular equivalence graph – is captured by a probability distribution of the form (3) over the space of equivalence graphs. A *measurement model*, that is, a conditional probability distribution $p(\mathbf{y}(t) \mid \mathbf{x}(t))$, captures what is known about sensing: If the current state $\mathbf{x}(t)$ is known, the current measurement $\mathbf{y}(t)$ is assumed to be independent of past and future states, and the measurement model is a unimodal version ($C = 1$) of a distribution of the form (3) on association graphs, with an equivalence graph forming the reference graph $\hat{\mathbf{x}}(t) = \Gamma_1$. Finally, the state variables are assumed to form a Markov sequence, so that their interdependence is entirely captured by the *transition model* $p(\mathbf{x}(t + \Delta t) \mid \mathbf{x}(t))$, again of the form (3), but on equivalence graphs only.

Given these models and an *initial estimate* $p(\mathbf{x}(0) \mid \mathbf{y}(0))$, Bayesian filtering then estimates for any desired time t the *posterior probability distribution* $p(\mathbf{X}(t) \mid \mathbf{Y}(t))$, where

$$\mathbf{X}(t) = \{\mathbf{x}(\tau) \mid 0 \leq \tau \leq t\} \quad \text{and} \quad \mathbf{Y}(t) = \{\mathbf{y}(\tau) \mid 0 \leq \tau \leq t\}$$

are the accumulated histories of states and measurements. This posterior distribution can be used, when desired, to obtain MAP point estimates of the identities associated with the observations:

$$\hat{\mathbf{X}}(t) = \arg\max_{\mathbf{X}(t)} p(\mathbf{X}(t) \mid \mathbf{Y}(t)) \; .$$

Bayesian filtering was outlined in the proposal for this grant, and is essentially unaffected by the changes of formulation described so far. Because of this, this method is summarized in Appendix A for completeness.

2.2 Batch Solution

The crucial computation in the proposed estimation procedure is initialization – or batch solution – that is, the computation of the posterior probability $p(\mathbf{X}(0) \mid \mathbf{Y}(0)) = p(\mathbf{x}(0) \mid \mathbf{y}(0))$. Here, $\mathbf{y}(0)$ is a set of association values p_{ij} between observations, and $\mathbf{x}(0)$ is a partition of the set of observations into distinct identities.

This computation can be viewed as a batch (that is, non-recursive) version of stochastic estimation, in which the probability distribution over possible identities is estimated from a fixed set of observations (association values). If the number of observations is relatively small, and all data is available ahead of time, initialization solves the whole data association problem. For data sets that grow indefinitely over time, propagation and update are necessary in addition, as outlined in Appendix A.

The posterior $p(\mathbf{x}(0) \mid \mathbf{y}(0))$ can be computed by first determining high-likelihood estimates $\hat{\mathbf{x}}(0)$ of $\mathbf{x}(0)$ from $\mathbf{y}(0)$. We can then use a Mallows-like expression of the form (3) with the estimates playing the role of the Γ_c equivalence graphs.

The computation of likely graph partitions $\hat{\mathbf{x}}(0)$ from a graph $\mathbf{y}(0)$ of association values has been viewed in the literature in different but related ways:

- A partition of the node set $V(t)$ can be represented by an *equivalence graph* made of disjoint cliques: Two nodes are equivalent if and only if they are in the same clique.

Identity management can then be cast as the problem of finding the equivalence graph closest to $G(t)$, in a metric to be specified. Formally, equivalence graphs are special cases of association graphs, namely, those made of disjoint cliques and with binary edge weights.

- The same partition of $V(t)$ can be viewed as a *clustering* of $V(t)$ in a metric specified partially by the association weights in $P(t)$. Inference management then optimizes a ratio between intra-cluster and inter-cluster spread, suitably defined.
- If the number k of sets in the clustering or partition of $V(t)$ were known, the necessary computation could be restated as a *graph-cut* problem: Find a minimum-weight subset of $E(t)$ whose removal separates the graph into k connected components. Inference management can then be phrased as a joint estimation of k and the corresponding optimal k-cut.

These approaches differ by what is being optimized: a distance between graphs; a spread ratio; or the cost of a cut. They correspond to different approaches and algorithms in the literature. However, all these algorithms are designed to work with – or imply – additive measures for graph cuts, in that they use the sum of association values along edges to determine whether a set of edges should be removed.

To experiment with these approaches, we developed a software infrastructure with the elements described in the next Section. Section 2.4 then describes a key problem common to approaches that estimate $\hat{\mathbf{x}}(0)$ by additive measures of graph cuts, and Section 2.5 shows our solution to this problem.

2.3 Experimental Setup

The main question addressed by our experiments is the extent to which the initial state estimate $p(\mathbf{X}(0) \mid \mathbf{Y}(0))$ – henceforth abbreviated to $p(\mathbf{X} \mid \mathbf{Y})$ – reflects ground truth, as determined by simulation, with varying amounts of uncertainty in the input association graph.

Answering this question required the programming of the following modules:

- A *simulation* module that creates a partition graph Γ and perturbs it in a controlled way to produce an association graph G. Perturbations include modification of the binary values on the edges of Γ into association values in the interval $[0, 1]$ for the edges of G. They also include removal of a specified number of edges from G, to simulate unspecified edges.
- The *SV* module itself, which takes an association graph G and an integer k as inputs and produces a set of edges of minimum weight whose removal leaves k connected components.
- A *clustering* module that takes an association graph G, fixed integers C and N with $N >> C$, and a real number r with $0 \leq r \leq 1$. This module generates N random values of k by a Dirichlet process. For each value of k, the clustering module draws a fraction r of edges from G, and runs a graph partition estimation algorithm on the resulting graph and with parameter k. This produces a sequence $\Gamma_1, \ldots, \Gamma_N$ of association graphs which are then clustered with the C-means algorithm. The output from this modules is the set of resulting cluster centers $\Gamma_1, \ldots, \Gamma_C$, together with the dispersion parameters $\sigma_1, \ldots, \sigma_C$ computed through equation (2).

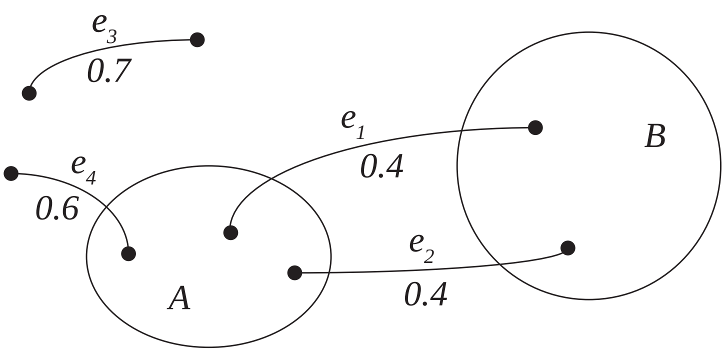

Figure 1: Algorithms that measure the cost of a cut by the sum of its edge weights will not remove the weak, spurious edges e_1, e_2 between components A and B of this graph until after "dangling" connections like e_3, e_4 are severed.

- An *evaluation* module that determines the likelihood of association graph G by computing expression (3).
- A *display* module that shows statistics of the results as a function of the structure of G and the extent of its perturbations. Likelihood is expected to be a decreasing function of the number of removed edges and the extent of perturbation of the remaining association values.

2.4 The Additivity Problem

After implementing several of them in Matlab, and testing them in simulation, we realized that this additivity leads to brittleness in the presence of spurious association values p_{ij}. Figure 1 illustrates this problem. In this Figure, two strong connected components A, B of a graph are tied to each other by two edges e_1, e_2 with relatively low association values. Any algorithm that is based on additive cut measures will end up preserving this weak connection at least until after "dangling" edges such as e_3, e_4 are severed.

The algorithm by Saran and Vazirani [8] to find a k-cut of graph that is a factor $2 - 2/k$ of minimal may serve as an illustration. This method requires specifying the number k of identities (eventual graph components), which is generally unknown. To address this difficulty, we embedded the k-cut computation within an estimator that models k as a Dirichlet process [7]: We draw from this process to hypothesize k, and we solve the corresponding k-cut problem, as illustrated in Section 2.3.

In our experiments, this algorithm needed large values of k in order to sever weak connections with a few edges (as exemplified by edges e_1, e_2 in Figure 1) if there exist several, somewhat stronger connections with individual, otherwise isolated nodes (as exemplified by edges e_3, e_4 in the Figure). At these high values of k, the likelihood that also unwanted cuts are made becomes large, and estimated partitions become meaningless. We call this issue the *additivity problem*.

2.5 The Partition Filtration

To address the additivity problem, we developed a method that is based on the concept of a *partition filtration*. This concept leads to straightforward estimates of high-likelihood cuts.

A partition Γ_2 is said to be a *refinement* of partition Γ_1 of the same set of nodes,

$$\Gamma_1 \prec \Gamma_2 \; ,$$

if every set of Γ_2 is contained in (or possibly coincident with) a set of Γ_1.

Given an association graph G with vertex set V and association values (edge weights) p_{ij} between nodes i and j, form the complete graph G by replacing missing edges with zero-weight edges. Then, if π is a real number between 0 and 1, form the partition $\Gamma(G, \pi)$ as the set of connected components of the binary graph with vertex set V and with an edge between nodes i and j if and only if

$$p_{ij} \geq \pi \; .$$

In other words, keep the edges of G that have weight at least π, and compute the corresponding connected components.

Then, the parametric family of graphs $\Gamma(G, \pi)$ is a *filtration*, in the sense that

$$0 \leq \pi_1 \leq \pi_2 \leq 1 \Rightarrow \Gamma(G, \pi_1) \prec \Gamma(G, \pi_2) \; .$$

In this filtration, $\Gamma(G, 0)$ is the complete graph on V and $\Gamma(G, 1)$ is the trivial graph where every node in V is a separate component (empty graph).

Then, we fix a real parameter ϵ between 0 and $1/2$, and define the ϵ-set of reference partitions $\Gamma_1, \ldots, \Gamma_C$ in equation (3) to be the set of all partitions in the family $\Gamma(G, \pi)$ such that

$$\frac{1}{2} - \epsilon \leq \pi \leq \frac{1}{2} + \epsilon \; .$$

With this definition, the computation of equation (3), the crucial component of the identity management problem, has become very efficient.

The parameter ϵ determines how far one is willing to go from the *standard partition* $\Gamma(G, 1/2)$, which essentially takes the association probabilities p_{ij} at their face values. Note that missing associations are treated as evidence of no association in this context, that is, they are made equivalent to $p_{ij} = 0$. This reflects the conservative stance whereby identities are equated only in the presence of positive evidence.

The first panel in Figure 2 shows a random association graph. Blue edges have weights significantly higher than 1/2, and red edges have weights significantly lower than 1/2. Thus, edges in this graph simulate observations that have good confidence, that is, whose association values are far from 1/2. The remaining panels in the figure show all the partitions in the complete filtration, obtained when $\epsilon = 1/2$. Each class in a partition is shown by its clique, and the range of π values for which each partition is valid is shown at the top of each graph. In terms of a partition distance [5], which measures how many nodes have to be moved between two partitions to make them equal, the partitions in the filtration are very different from each other. In terms of the compatibility function defined in equation 1, on the other hand, the partitions for small values of ϵ are very compatible with each other, because they differ by a small number of edges. Thus, filtration partitions and compatibility work well together in the definition (3) for the probability distribution over association graphs. First, graphs that differ by a small number of edges are mutually compatible. Second, the partitions in a filtration for a given association graph provide a set of reference graphs that account well for uncertainty in the association values. Third, the extent of the range of π for a partition can be used as a measure of the *persistence* of that partition to changes in the threshold π. For instance, the partition with highest persistence in Figure 2 is the left graph in the third row (boxed), with a value of persistence equal to $0.87 - 0.27 = 0.6$.

In other words, the partition filtration is an efficient solution to the batch version of identity management, and can therefore be used for initialization in the online version.

References

[1] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[2] G. S. Fishman. *Monte Carlo*. Springer-Verlag, 2003.

[3] John Geweke. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339, 1989.

[4] L. J. Guibas. The identity management problem – a short survey. In *Proc. 11th International Conference on Information Fusion*, pages 1–7, 2008.

[5] D. A. Konovalov, B. Litow, and N. Bajema. Partition-distance via the assignment problem. *Bioinformatics*, 21(10):2463–2468, 2005.

[6] C. L. Mallows. Non-null ranking models I. *Biometrika*, 44:114–130, 1957.

[7] C. Rasmussen. The infinite gaussian mixture model. *Advances in Neural Information Processing Systems (NIPS)*, 12:554–560, 2000.

[8] H. Saran and V. V. Vazirani. Finding k-cuts within twice the optimal. *SIAM Journal on Computing*, 24(1):101–108, 1995.

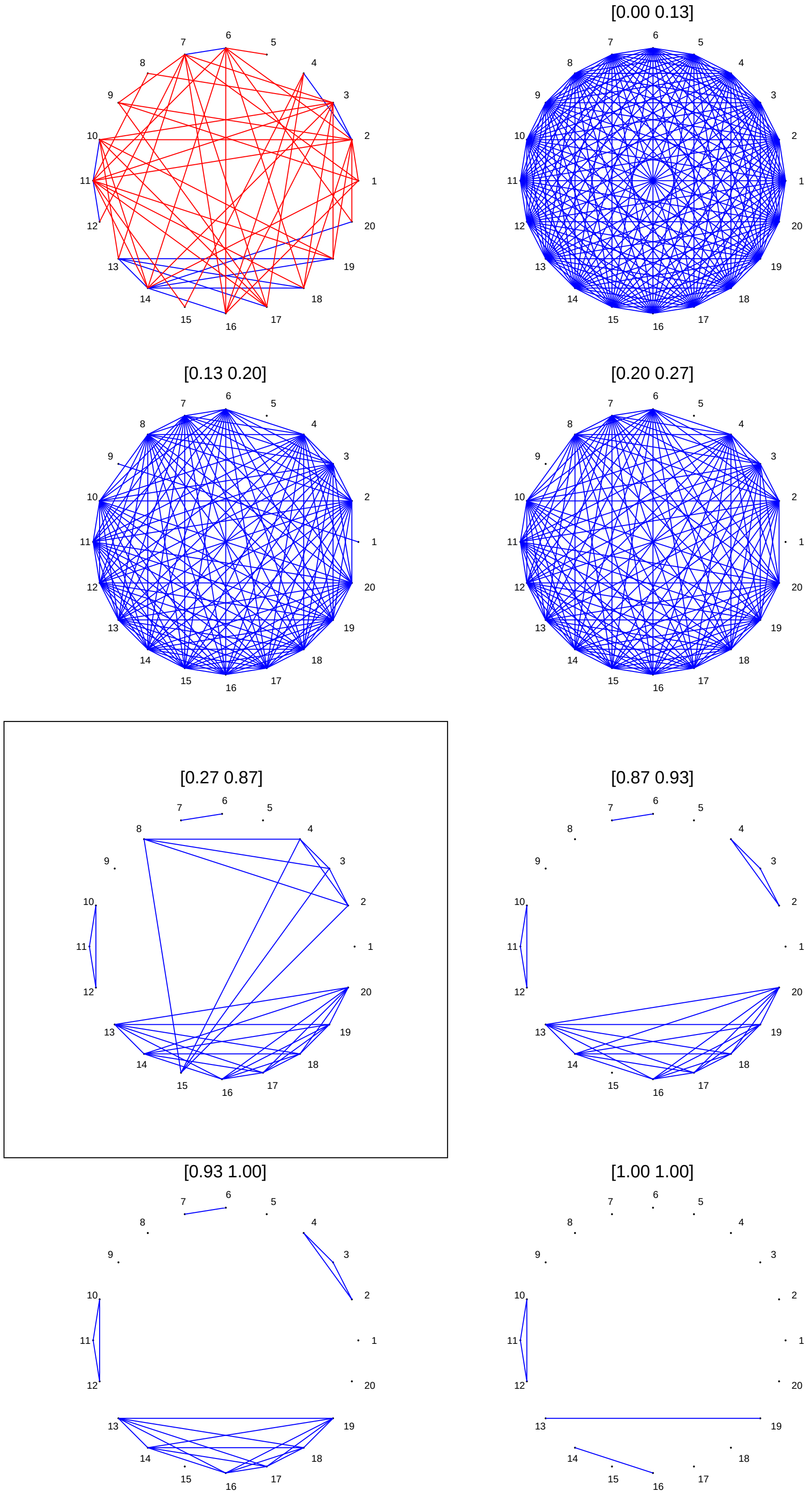

Figure 2: Top left: An observation graph. Blue edges correspond to strong associations, red edges to weak ones. Other panels: Partitions in the complete filtration for the graph at top left. The ranges of π that result in each partition are shown above each partition. The boxed partition is the most persistent one.

A Bayesian Filtering

Simple manipulation [3] shows that the following recursive formula holds for the posterior probability distribution $p(\mathbf{X}(t) \mid \mathbf{Y}(t))$:

$$p(\mathbf{X}(t{+}\Delta t) \mid \mathbf{Y}(t{+}\Delta t)) = p(\mathbf{X}(t) \mid \mathbf{Y}(t)) \, \frac{p(\mathbf{y}(t+\Delta t) \mid \mathbf{x}(t+\Delta t)) \, p(\mathbf{x}(t+\Delta t) \mid \mathbf{x}(t))}{p(\mathbf{y}(t+\Delta t) \mid \mathbf{Y}(t))} \; . \tag{4}$$

On the right-hand side, $p(\mathbf{X}(t) \mid \mathbf{Y}(t))$ is the old posterior, for which an initial value is known ($p(\mathbf{X}(0) \mid \mathbf{Y}(0))$), while $p(\mathbf{y}(t+\Delta t) \mid \mathbf{x}(t+\Delta t))$ is the measurement model and $p(\mathbf{x}(t+\Delta t) \mid \mathbf{x}(t))$ is the transition model.

The remaining normalization term, $p(\mathbf{y}(t+\Delta t) \mid \mathbf{Y}(t))$, is in general difficult to compute analytically. This suggests [3] using a sampling approach, because Monte Carlo sampling does not require knowing the normalization factor. Instead of computing the right-hand side of equation (4), one can thus sample from it, thereby representing the distribution on the left-hand side though a collection of samples. This approach in turn requires the ability to draw samples from the old posterior $p(\mathbf{X}(t) \mid \mathbf{Y}(t))$, and to evaluate both transition model and measurement model $p(\mathbf{y}(t) \mid \mathbf{x}(t))$ pointwise.

Sampling from the old posterior $p(\mathbf{X}(t) \mid \mathbf{Y}(t))$ directly is also difficult. Recursive importance sampling [1] circumvents this difficulty by sampling instead from a simpler distribution $\pi(\mathbf{X}(t) \mid \mathbf{Y}(t))$, called the *importance function*, whose support includes that of the old posterior, and which is designed to have the following form to enable recursive computation:

$$\pi(\mathbf{X}(t) \mid \mathbf{Y}(t)) = \pi(\mathbf{x}(0) \mid \mathbf{y}(0)) \prod_{t_k \le t} \pi(\mathbf{x}(t_k) \mid \mathbf{x}(t_0), \ldots, \mathbf{x}(t_{k-1}), \mathbf{Y}(t_k)) \; . \tag{5}$$

It can then be shown [1] that the following *Sequential Importance Sampling* (SIS) framework yields posterior weighted samples $x_k^{(i)}$ with weights $w_k^{(i)}$ at time t_k for $k = 0, 1, \ldots$:

- For $i = 1, \ldots, N$, sample $\mathbf{x}_k^{(i)} \sim \pi(\mathbf{x} \mid \mathbf{x}_0^{(i)}, \ldots, \mathbf{x}_{k-1}^{(i)}, \mathbf{Y}(t_k))$.
- For $i = 1, \ldots, N$, evaluate the unnormalized importance weights:

$$u_k^{(i)} = u_{k-1}^{(i)} \, \frac{p(\mathbf{y}(t_k) \mid \mathbf{x}_k^{(i)}) \, p(\mathbf{x}_k^{(i)} \mid \mathbf{x}_{k-1}^{(i)})}{\pi(\mathbf{x}_k^{(i)} \mid \mathbf{x}_0^{(i)}, \ldots, \mathbf{x}_{k-1}^{(i)}, \mathbf{Y}(t_k))}$$

- For $i = 1, \ldots, N$, normalize the importance weights:

$$w_k^{(i)} = \frac{u_k^{(i)}}{\sum_{j=1}^{N} u_k^{(j)}} \; .$$

Because of the mixture form (3) of the conditional probability density $p(\mathbf{x} \mid \mathbf{y})$, initialization amounts to estimating the number C of mixture components, the C reference equivalence graphs in the vector $\boldsymbol{\gamma} = (\Gamma_1, \ldots, \Gamma_C)^T$, and the C dispersion parameters in the vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_C)^T$.

Sequential importance sampling requires defining an importance function of the form (5) from which is it easy to sample. While the requirements on this function are very mild as far as statistical convergence is concerned, efficiency demands that the function approximates the true posterior well. In the context of identity management, this means that the term

$$\pi(\mathbf{x}(t_k) \mid \mathbf{x}(t_0), \ldots, \mathbf{x}(t_{k-1}), \mathbf{Y}(t_k))$$

that appears in equation (5) be a plausible probabilistic description of what new edges appear in the equivalence graph that describes estimated identities as new observations become available.

We propose to characterize this distribution by a random walk in the space of equivalence graphs. A walk must move from graph to graph, and steps are easily generated: the start equivalence graph is merely a partition of its nodes, and a new partition can be generated by moving random nodes between random classes of the partition.

In order to condition the resulting walk on the observations $\mathbf{y}(t)$, we use the Metropolis idea [2]: evaluate the likelihood ratio $r = p(\omega')/p(\omega)$, where ω is the old partition and ω' is the new one. If r is greater than one, accept the new step. Otherwise, accept it with probability r, and reject it with probability $1 - r$.

So far, it has been assumed that the set $\mathbf{Y}(t)$ of measurements grows over time, and the state $\mathbf{X}(t)$ grows with it. An efficient recursive procedure, on the other hand, relies on $\mathbf{X}(t)$ being bounded in size. This is achieved by forgetting old observations. “Old” here can be measured in a principled fashion by referring to two related probability distributions. The first is the probability $p_{j|i}$ that a target observed at sensor i subsequently appears at sensor j first. The second is the conditional probability $\pi_{ij}(t)$ that such a target arrives at j after a time interval t. Then, given a small probability ϵ, the observation $x_i(t_i)$ is removed if

$$\max_{p_{j|i} > \epsilon} \int_{\tau \geq t} \frac{\pi_{ij}(\tau)}{p_{j|i}} \, d\tau < \epsilon \; .$$

In words, an observation is forgotten when it is so old that the probability that its target has not yet reappeared elsewhere is negligible.